High finite-sample efficiency and robustness based on distance-constrained maximum likelihood

Ricardo Maronna and Víctor Yohai

Ricardo A. Maronna is Consulting Professor, Mathematics Department, National University of La Plata, C.C. 172, La Plata 1900, Argentina (E-mail: rmaronna@retina.ar. Víctor J. Yohai is Professor Emeritus, Mathematics Department, Faculty of Exact Sciences, Ciudad Universitaria, 1428 Buenos Aires, Argentina (E-mail: victoryohai@gmail.com). This research was partially supported by Grants W276 from Universidad of Buenos Aires, PIP's 112-2008-01-00216 and 112-2011-01- 00339 from CONICET and PICT 2011-0397 from ANPCYT, Argentina.

ABSTRACT

Good robust estimators can be tuned to combine a high breakdown point and a specified asymptotic efficiency at a central model. This happens in regression with MM- and τ-estimators among others. However, the finite-sample efficiency of these estimators can be much lower than the asymptotic one. To overcome this drawback, an approach is proposed for parametric models, which is based on a distance between parameters. Given a robust estimator, the proposed one is obtained by maximizing the likelihood under the constraint that the distance is less than a given threshold. For the linear model with normal errors and using the MM estimator and the distance induced by the Kullback-Leibler divergence, simulations show that the proposed estimator attains a finite-sample efficiency close to one, while its maximum mean squared error is smaller than that of the MM estimator. The same approach also shows good results in the estimation of multivariate location and scatter.

Key words: Finite-sample efficiency; contamination bias; MM estimators, S estimators.

1 Introduction

Since the seminal work of Huber (1964) and Hampel (1971), one of the main concerns of the research in robust statistics has been to derive statistical procedures that are simultaneously highly robust and highly efficient under the assumed model. The efficiency of an estimator is usually measured by the asymptotic efficiency, that is, by the ratio between the asymptotic variances of the maximum likelihood estimator (henceforth MLE) and of the robust estimator. However if the sample size n is not very large, this asymptotic efficiency may be quite different from the finite sample size one, defined as the ratio between the mean squared errors (MSE) of the MLE and of the robust estimator, for samples of size n. However, it is obvious that for practical purposes only the finite sample size efficiency matters.

Consider for example the case of a linear model with normal errors. In this case the MLE of the regression coefficients is the least squares estimator (LSE). It is well known that this estimator is very sensitive to outliers, and in particular its breakdown point is zero. To overcome this problem, several estimators combining high asymptotic breakdown point and high efficiency have been proposed. Yohai (1987) proposed MM-estimators, which have 50% breakdown point and asymptotic efficiency as close to one as desired. Yohai and Zamar (1988) proposed $\tau-$estimates, which combine the same two properties as MM-estimators. Gervini and Yohai (2002) proposed regression estimators which simultaneously have 50% breakdown point and asymptotic efficiency equal to one.

However, as will be seen in Section 2.1, when n is not very large the finite sample efficiency of these estimators may be much smaller than the asymptotic one. On the other hand, a 50% breakdown point does not guarantee that the estimator is highly robust. In fact, this only guarantees that given $\varepsilon < 0.5$ there exists $K(\varepsilon)$ such that if the data are contaminated with a fraction of outliers smaller than ε, the norm of the difference between the estimator and the true value is smaller than $K(\varepsilon)$. However $K(\varepsilon)$ may be very large, which makes the estimator unstable under outlier contamination of size ε. Bondell and Stefanski (2013) proposed a regression estimator with maximum breakdown point and high finite-sample efficiency. However, as it will be seen in Section 2.1, the price for this efficiency is a serious loss of robustness.

The purpose of this paper is to present estimators which have a high finite sample size efficiency and robustness even for small n. Besides, these estimators are highly robust using a robustness criterion better than the breakdown point, namely, the maximum MSE for a given contamination rate ε.

The procedure to define the proposed estimators is very general and may be applied to any parametric or semiparametric model. However in this paper the details are given only to estimate the regression coefficients in a linear model and the multivariate location and scatter of a random vector.

To define the proposed estimators we need an initial robust estimator, not necessarily with high finite sample efficiency. Then the estimators are defined by maximizing the likelihood function subject to the estimate being sufficiently close to the initial one. Doing so we can expect that the resulting estimator will have the maximum possible finite sample efficiency under the assumed model compatible with proximity to the initial robust estimator. This proximity guarantees the robustness of the new estimator.

The formulation of our proposal is as follows. Let D be a distance or discrepancy measure between densities. As a general notation, given a family of distributions with observation vector $\mathbf{z}$, parameter vector θ and density $f(\mathbf{z},\theta)$, put $d(\theta_1,\theta_2) = D(f(\mathbf{z},\theta_1), f(\mathbf{z},\theta_2))$. Let $\mathbf{z}_i$, $i = 1,..,n$ be i.i.d. observations with distribution $f(\mathbf{z},\theta)$, and let $\widehat{\theta}_0$ be an initial robust estimator. Call $L(\mathbf{z}_1,...,\mathbf{z}_n;\theta)$ the likelihood function. Then our proposal is to define an estimator $\widehat{\theta}$ as

$$\widehat{\theta} = \arg\max_{\theta} L(\mathbf{z}_1,...,\mathbf{z}_n;\theta) \quad \text{with } d\left(\widehat{\theta}_0,\theta\right) \leq \delta \tag{1}$$

where δ is an adequately chosen constant that may depend on n. We shall call this proposal “distance-constrained maximum likelihood’ (DCML for short).

Several dissimilarity measures, such as the Hellinger distance, may be employed for this purpose. We shall employ as D the Kullback-Leibler (KL) divergence, because, as it will be seen, it yields easily manageable results. Therefore the d in

(1) will be

$$d_{\mathrm{KL}}\left(\theta_1,\theta_2\right)=\int_{-\infty}^{\infty}\log\left(\frac{f\left(\mathbf{z},\theta_1\right)}{f\left(\mathbf{z},\theta_2\right)}\right)f\left(\mathbf{z},\theta_1\right)d\mathbf{z}.$$

In Sections 2 and 3 we apply this procedure to the linear model and to the estimation of multivariate location and scatter, respectively. In Section 4 we apply the DCML estimator to real data.

2 Regression

Consider the family of distributions with $\mathbf{z}=(\mathbf{x},y)$, with $\mathbf{x}\in R^p$ and $y\in R$, satisfying the model $y=\mathbf{x}'\beta+\sigma u$, where $u\sim\mathrm{N}(0,1)$ is independent of $\mathbf{x}\in R^p$. Here $\theta=(\beta,\sigma)$. Let $\widehat{\theta}_0=\left(\widehat{\beta}_0,\widehat{\sigma}_0\right)$ be an initial robust estimator of regression and scale. We will actually consider σ as a nuisance parameter, and therefore we have

$$d_{\mathrm{KL}}\left(\beta_0,\beta\right)=\frac{1}{\sigma^2}\left(\beta-\beta_0\right)'\mathbf{C}\left(\beta-\beta_0\right) \tag{2}$$

with $\mathbf{C}=\mathrm{E}\mathbf{x}\mathbf{x}'$.

Here we replace σ with its estimator $\widehat{\sigma}_0$. The natural estimator of $\mathbf{C}$ would be $\widehat{\mathbf{C}}=n^{-1}\mathbf{X}'\mathbf{X}$, where $\mathbf{X}$ is the $n\times p$ matrix with rows $\mathbf{x}_i'$. Since it is not robust, we will employ a robust version thereof. Put for $\beta\in R^p$ $r_i(\beta)=y_i-\mathbf{x}'\beta$, the residuals from β. Most "smooth" robust regression estimators, like S-estimators (Rousseeuw and Yohai, 1984) and MM-estimators satisfy the estimating equations of an M-estimator, which can be written as weighted normal equations, namely

$$\sum_{i=1}^{n}W\left(\frac{r_i(\beta)}{\widehat{\sigma}_0}\right)\mathbf{x}_i r_i\left(\beta\right)=\mathbf{0}, \tag{3}$$

where W is a "weight function". Then we define, as in (Yohai et al., 1991)

$$\mathbf{C}_w=\frac{1}{\sum_{i=1}^{n}w_i}\sum_{i=1}^{n}w_i\mathbf{x}_i\mathbf{x}_i, \tag{4}$$

with $w_i = W\left(r_i(\widehat{\beta}_0)/\widehat{\sigma}_0\right)$.

It is immediate that (1) is equivalent to minimizing $\sum_{i=1}^{n} r_i(\beta)^2$ subject to $d_{\text{KL}}\left(\widehat{\beta}_0, \beta\right) \leq \delta$. Call $\widehat{\beta}_{\text{LS}}$ the LSE. Put for a general matrix $\mathbf{V}$:

$$d_{\mathbf{V}} = \frac{1}{\widehat{\sigma}_0^2}\left(\widehat{\beta}_0 - \widehat{\beta}_{\text{LS}}\right)\mathbf{V}\left(\widehat{\beta}_0 - \widehat{\beta}_{\text{LS}}\right).$$

Then using Lagrange multipliers, a straightforward calculation shows that

$$\widehat{\beta} = \begin{cases} \widehat{\beta}_{\text{LS}} & \text{if} \quad d_{\mathbf{C}_w} \leq \delta \\ \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}_w\right)^{-1}\left(\mathbf{X}'\mathbf{X}\widehat{\beta}_{\text{LS}} + \lambda\mathbf{C}_w\widehat{\beta}_0\right) & \text{else} \end{cases}, \tag{5}$$

where λ is determined from the equation $d_{\text{KL}}\left(\widehat{\beta}_0, \beta\right) = \delta$ and $\mathbf{C}_w$ is defined in (4). We thus see that $\widehat{\beta}$ is a linear combination of $\widehat{\beta}_0$ and $\widehat{\beta}_{\text{LS}}$.

Another approach is as follows. Define $\widehat{\beta}$ as the minimizer of $\sum_{i=1}^{n} r_i(\beta)^2$ subject to $d_{\widehat{\mathbf{C}}} \leq \delta$. In this case the solution is explicit:

$$\widehat{\beta} = t\widehat{\beta}_{\text{LS}} + (1-t)\widehat{\beta}_0, \tag{6}$$

where $t = \min\left(1, \sqrt{\delta/d_{\widehat{\mathbf{C}}}}\right)$. Since $d_{\widehat{\mathbf{C}}}$ is not robust, we now replace it with $d_{\mathbf{C}_w}$, and therefore we choose

$$t = \min\left(1, \sqrt{\frac{\delta}{d_{\mathbf{C}_w}}}\right). \tag{7}$$

The difference between both versions (5) and (6) showed to be negligible for all practical purposes.

It is easy to show that if $\widehat{\beta}_0$ is regression- and affine-equivariant, so is $\widehat{\beta}$.

2.1 Simulations

We now consider the model

$$y_i = \mathbf{x}_i'\beta + \sigma u_i, \; i = 1, ..., n, \tag{8}$$

with $\beta \in R^p$ and $u_i \sim \mathrm{N}(0,1)$ independent of $\mathbf{x}_i$. The performance of each estimator $\widehat{\beta}$ will be measured by its prediction squared error, which is equivalent to $\left(\widehat{\beta}-\beta\right)'\mathbf{C}_x\left(\widehat{\beta}-\beta\right)$,where $\mathbf{C}_x$=E$\mathbf{xx}'$. Since all estimators considered are regression-equivariant, there is no loss of generality in taking $\beta = \mathbf{0}$. In all cases, the distributions are normalized so that $\mathbf{C}_x = \mathbf{I}$, and therefore the criterion will be simply $\left\|\widehat{\beta}\right\|^2$ where $\|.\|$ stands for the Euclidean norm.

As initial estimator $\widehat{\beta}_0$ we chose the MM estimator with 85% asymptotic efficiency and bisquare $\rho-$function:

$$\rho_{\mathrm{bis}}(d) = 1 - \mathrm{I}(d \leq 1)(1-d)^3, \tag{9}$$

where $\mathrm{I}(.)$ denotes the indicator function. The MM estimator needs a starting regression estimator and a starting scale, which were supplied by the S estimator $\widehat{\beta}_{\mathrm{SE}}$ with the same ρ.

The reason for choosing 85% efficiency is that the maximum bias of the resulting estimator for normal predictors is the same as that of the regression S-estimator, as explained in Section 5.9 of (Maronna et al., 2006).

An S-estimator was also considered as an initial estimator. However, the asymptotic efficiency of these estimators is known to be less than 33%, and the finite-sample efficiency is still lower. Therefore to attain acceptable efficiencies for DCML the values δ should have to be substantially larger than the ones we employed (given in (11) below), which would entail a serious loss in robustness. These assertions were confirmed by the simulations and therefore MM was the estimator of choice.

The initial scale $\widehat{\sigma}_0$ is a scale M estimator of the residuals, defined as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho_{\mathrm{bis}}\left(\frac{y_i - \mathbf{x}_i'\widehat{\beta}_0}{\mathbf{c}_0.\widehat{\sigma}_0}\right) = \gamma, \tag{10}$$

where $c_0 = 1.547$ makes $\widehat{\sigma}_0$ consistent in the normal case and $\gamma = 0.5\,(1-p/n)$

The constant δ in (1) is chosen as

$$\delta_{p,n} = 0.3\frac{p}{n}. \tag{11}$$

To justify (11) note that under the model the distribution of $nd_{\mathrm{KL}}\left(\widehat{\beta}_0, \widehat{\beta}_{\mathrm{LS}}\right)$ is approximately that of vz where $z \sim \chi^2_p$ and v is some constant, which implies that $\mathrm{E}d_{\mathrm{KL}}\left(\widehat{\beta}_0, \widehat{\beta}_{\mathrm{LS}}\right) \approx vp/n$. Therefore in order to control the efficiency of $\widehat{\beta}$ it seems reasonable to take δ of the form cp/n for some c. The value $c = 0.3$ was arrived at after exploratory simulations aimed at striking a balance between efficiency and robustness.

2.1.1 Scenarios

Since the results may depend on the distribution of the predictors, we considered five cases, all of them including an intercept. Here each predictor vector has the form $\mathbf{x} = (1, x_1, .., x_p)'$, where the x_js are i.i.d. random variables with distribution F. Note that here the number of parameters is $p + 1$. In the first three cases F is standard normal, uniform in [0,1] (short-tailed) and Student with four degrees of freedom (moderately heavy-tailed). In the other two, the x_js are the *squares* of standard normal and uniform variables. The Student distribution was excluded for in this case $\mathbf{C}_x$=E$\mathbf{xx'}$ does not exist since it involves the fourth moments of the t_4 distribution.

We took $p = 5$, 10 and 20, and $n = Kp$ with $K = 5$, 10 and 20.

For each n and p we first computed the finite sample efficiency. Then to assess the estimators' robustness we contaminated the data as follows. For a contamination rate $\varepsilon \in (0, 1)$ let $m = [n\varepsilon]$ where [.] stands for the integer part. Then for $i \leq n - m$, $(\mathbf{x}_i, y_i)$ were generated according to model (8), and for $i > n - m$ we put $\mathbf{x}_i = (1, x_0, 0, ..., 0)'$ and $y_i = x_0K$, where the parameter K which regulates the slope of the contamination took on a range of values in order to determine the worst possible situations. The effect of the contamination would be to drag the first slope towards K. We took $x_0 = 5$ and K ranging between 0.5 and 2 with intervals of 0.1. We employed $\varepsilon = 0.1$ and 0.2. The number of replications was $N_{\text{rep}} = 1000$ and 200 for the uncontaminated and contaminated cases respectively.

For a given scenario and estimator $\widehat{\beta}$ call $\widehat{\beta}_k$, $k = 1, ..., N_{\text{rep}}$ the Monte Carlo values. As measure of performance we employed the mean squared error: $\text{MSE} = \text{ave}_k \left\{ \left\| \widehat{\beta}_k \right\|^2 \right\}$ where "ave" stands for the average.

2.1.2 Estimators

The estimators considered were: the Least Squares estimator, the regression S-estimator with bisquare scale (S-E), the MM estimator with bisquare loss function and 85% asymptotic efficiency, the Gervini-Yohai (2002) estimator (G-Y), the Bondell-Stefanski (2013) estimator (B-S), and the proposed estimator (DCML). Both versions (5) and (6) were considered, but since the latter yielded in general slightly better results, this is the one that is reported here. S-E, MM and G-Y were computed using the function lmRob of the R robust package. The code for B-S was kindly supported by the authors.

2.1.3 Efficiency

We deal first with the efficiencies. In order to synthesize the results, for each combination (p, n) we took for each estimator the minimum efficiencies under normal errors over the five distributions, with respect to the MLE. The results are displayed in Table 1.

p	n	S-E	MM	G-Y	B-S	DCML
5	25	0.306	0.652	0.657	0.952	0.843
	50	0.270	0.773	0.799	0.990	0.944
	100	0.261	0.810	0.860	0.996	0.981
10	50	0.276	0.686	0.702	0.986	0.917
	100	0.276	0.777	0.821	0.997	0.977
	200	0.250	0.808	0.893	0.999	0.990
20	100	0.289	0.699	0.723	0.996	0.948
	200	0.254	0.774	0.841	0.999	0.984
	400	0.242	0.820	0.913	0.999	0.998

Table 1: Minimum efficiencies of estimators for normal errors over all x distributions

We note the following:

- The efficiency of S-E is low, as can be expected
- When n/p is "small", the worst finite-sample efficiency of MM can be much lower than its nominal asymptotic one of 85%. The worst cases with $n/p = 5$ corresponded to normal $\mathbf{x}_i$ with a quadratic term.
- The worst efficiency of G-Y is also low for small n/p.
- DCML outperforms both its initial estimator MM and G-Y.
- B-S shows the highest efficiencies in all cases.

Table 2 shows the efficiencies of the estimators with respect to the MLE for model (8) with Student errors u_i with 3 and 5 degrees of freedom ("d.f.").

df	p	n	S-E	MM	G-Y	B-S	DCML
3	5	25	0.453	0.828	0.799	0.875	0.893
		50	0.443	0.917	0.859	0.883	0.912
		100	0.477	0.949	0.870	0.871	0.900
	10	50	0.400	0.857	0.826	0.883	0.897
		100	0.418	0.928	0.865	0.892	0.917
		200	0.447	0.941	0.861	0.890	0.901
	20	100	0.424	0.880	0.854	0.904	0.943
		200	0.413	0.934	0.881	0.886	0.904
		400	0.447	0.946	0.867	0.863	0.883
5	5	25	0.384	0.747	0.733	0.934	0.896
		50	0.391	0.921	0.886	0.916	0.940
		100	0.398	0.919	0.875	0.925	0.946
	10	50	0.351	0.796	0.782	0.946	0.948
		100	0.350	0.894	0.878	0.933	0.946
		200	0.374	0.931	0.904	0.928	0.940
	20	100	0.368	0.828	0.821	0.940	0.966
		200	0.349	0.900	0.883	0.927	0.947
		400	0.371	0.923	0.898	0.936	0.935

Table 2: Efficiencies of estimators for Student errors with 3 and 5 degrees of freedom, and normal predictors

Here MM, G-Y, B-S and DCML exhibit high efficiencies, and none clearly dominates the others.

2.1.4 Robustness

We begin with the results of a typical case, Figure 1 displays the MSEs of the estimators for $p = 10$, $n = 200$, normal $\mathbf{x}$, and $\varepsilon = 0.1$, for different values of the outlier size K.

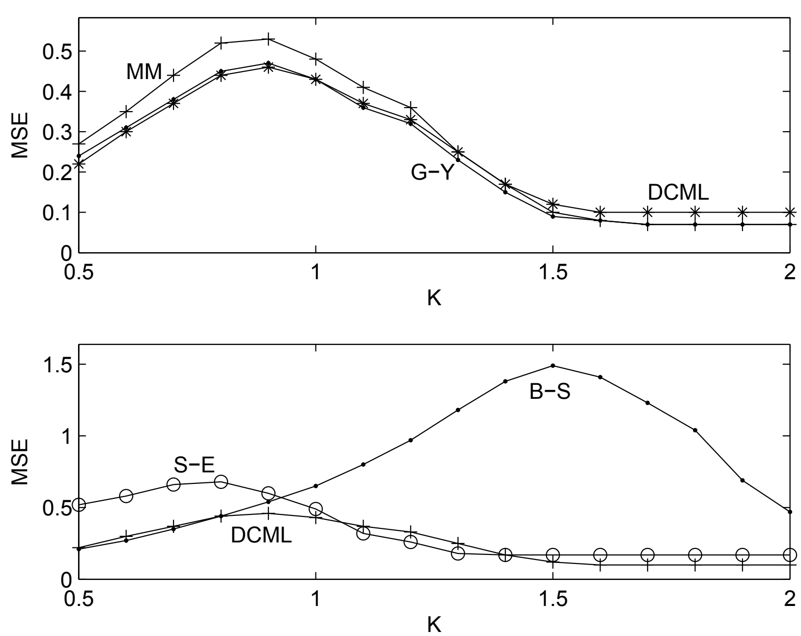

Figure 1: MSEs of regression estimators as a function of outlier size K for normal $\mathbf{x}$, $p = 10$, $n = 200$ and $\varepsilon = 0.1$.

In the upper panel it is seen that G-Y and DCML have similar behaviors, and that their maximum MSEs are smaller than that of MM. The lower panel shows that the MSEs of S-E and B-S are generally larger than that of DCML, the one of B-S being remarkably high.

Since all cases show approximately this same pattern, we display only the maximum MSEs over K. for normal $\mathbf{x}$. Table 3 shows the results.

.

Some comments are in order:

- The MSEs of G-Y and DCML are similar, the latter being lower in most cases. Both outperform MM, which in turn outperforms S-E.
- The price for the high efficiency of B-S is a high contamination bias.
- When $\varepsilon = 0.2$ and $n/p = 5$ all estimators have a remarkably high MSE.

As a closing comment, the joint consideration of Tables 1, 2 and 3 suggests that DCML shows the best balance between efficiency and robustness.

2.2 Asymptotic results

Assume $y = \mathbf{x}'\beta + u$, where u is independent of $\mathbf{x}$ and has distribution F. Call σ_0 be the limit value of the M-scale applied to u and $\mathbf{C} = \mathrm{E}(\mathbf{x}\mathbf{x}')$. It is well known that under general conditions the following expansions hold for the MM-estimator $\widehat{\beta}_0$ and the LS estimator $\widehat{\beta}_{\mathrm{LS}}$.

$$n^{1/2}(\widehat{\beta}_0 - \beta) = \frac{\sigma_0}{n^{1/2}\mathrm{E}\psi'(u_i/\sigma_0)} \sum_{i=1}^{n} C^{-1}\psi(\frac{u_i}{\sigma_0})\mathbf{x}_i + o\left(\frac{1}{n^{1/2}}\right),$$

ε	p	n	S-E	MM	G-Y	B-S	DCML
0.1	5	25	1.640	0.996	0.951	1.882	0.840
		50	1.143	0.692	0.637	1.557	0.590
		100	0.831	0.481	0.431	1.454	0.413
	10	50	2.730	1.588	1.514	2.602	1.268
		100	1.419	0.706	0.644	1.690	0.597
		200	0.973	0.543	0.475	1.530	0.463
	20	100	2.058	1.236	1.172	2.892	0.922
		200	1.212	0.633	0.569	1.940	0.515
		400	0.850	0.456	0.394	1.676	0.388
0.2	5	25	10.51	8.63	8.49	25.67	7.30
		50	5.24	3.79	3.70	9.58	3.34
		100	3.17	2.23	2.11	7.12	2.03
	10	50	14.23	12.00	11.86	23.37	9.99
		100	6.08	4.10	3.93	10.84	3.60
		200	3.55	2.47	2.32	8.70	2.27
	20	100	6.21	5.25	5.18	27.42	4.29
		200	3.52	2.70	2.60	11.94	2.35
		400	2.84	2.00	1.88	9.39	1.83

Table 3: Maximum mean squared errors of estimators with normal predictors for contaminated data

and

$$n^{1/2}(\widehat{\beta}_{\text{LS}}-\beta) = \frac{1}{n^{1/2}}\sum_{i=1}^{n} C^{-1}u_i\mathbf{x}_i + o\left(\frac{1}{n^{1/2}}\right)$$

It then follows from the Central Limit Theorem that the joint asymptotic distribution $J_{\mathbf{C},\mathbf{V}}$ of $n^{1/2}(\widehat{\beta}_{\text{LS}}-\beta, \widehat{\beta}-\beta_0)$ is $J_{\mathbf{C},\mathbf{V}} = \mathrm{N}_{2p}(\mathbf{0}, \mathbf{V}\otimes\mathbf{C}^{-1})$ where $\mathbf{V}=[V_{ij}]$ is a symmetric 2×2 matrix with elements

$$V_{11} = \mathrm{E}(u^2), \quad V_{12} = V_{21} = \sigma_0\frac{\mathrm{E}\left(u\psi\left(u/\sigma_0\right)\right)}{\mathrm{E}\left(\psi^{\prime}\left(u/\sigma_0\right)\right)}, \; V_{22} = \sigma_0^2\frac{\mathrm{E}\left(\psi^2\left(u/\sigma_0\right)\right)}{\mathrm{E}\left(\psi^{\prime}\left(u/\sigma_0\right)\right)} \tag{12}$$

Let $(\mathbf{z}_1,\mathbf{z}_2)^{\prime} \in R^{2p}$ be a random vector with distribution $J_{\mathbf{C},\mathbf{V}}$ and define

$$\mathbf{z}_3 = t\mathbf{z}_1 + (1-t)\mathbf{z}_2 \;\; \text{with} t = \min\left(1, \frac{0.3p}{(\mathbf{z}_2-\mathbf{z}_1)^{\prime}\mathbf{C}(\mathbf{z}_2-\mathbf{z}_1)}\right). \tag{13}$$

Then the distribution $H_{\mathbf{C},\mathbf{V}}$ of $\mathbf{z}_3$ is the same as the asymptotic distribution of $n^{1/2}(\widehat{\beta}-\beta)$. Note that since $\mathbf{z}_3$ is a nonlinear function of $(\mathbf{z}_1,\mathbf{z}_2)$, H is not necessarily normal. The following Theorem will be useful determine the distribution of $n^{1/2}\mathbf{b}^{\prime}(\widehat{\beta}-\beta)$ for any $\mathbf{b}\in\mathbf{R}^p$

Theorem 1 *If* $\mathbf{C}=\mathbf{I}$, *then the distribution of* $v=\mathbf{d}^{\prime}\mathbf{z}_3$ *is the same for any* $\mathbf{d}\in R^p$ *with* $||\mathbf{d}||=1$.

Proof: Let $\mathbf{D}$ be an orthogonal matrix with first row equal to $\mathbf{d}^{\prime}$ and let $\mathbf{v}_j=\mathbf{D}\mathbf{z}_j$, $1\leq j\leq 3$, where the $\mathbf{z}_j$s are defined above. It is easy to check that $(\mathbf{v}_1, v_2)$ has the same distribution as $(\mathbf{z}_1,\mathbf{z}_2)$, and that $\mathbf{v}_3$ satisfies

$$\mathbf{v}_3 = t\mathbf{v}_1 + (1-t)\mathbf{v}_2.$$

Besides, we have

$$(\mathbf{z}_2-\mathbf{z}_1)^{\prime}\mathbf{C}(\mathbf{z}_2-\mathbf{z}_1) = (\mathbf{v}_2-\mathbf{v}_1)^{\prime}\mathbf{C}(\mathbf{v}_2-\mathbf{v}_1)$$

and therefore

$$t = \min\left(1, \frac{0.3p}{(\mathbf{v}_2 - \mathbf{v}_1)'\mathbf{C}(\mathbf{v}_2 - \mathbf{v}_1)}\right)$$

Then $\mathbf{v}_3$ has the same distribution as $\mathbf{z}_3$, and therefore $v_{3,1} = \mathbf{d}'\mathbf{z}_3$ has the same distribution as $z_{3,1}$ independently of $\mathbf{d}$.$\blacksquare$

Call $G_{\mathbf{V}}(z)$ the distribution function of $v_{3,1}$.Suppose now that we want the distribution of $w = \mathbf{b}'\mathbf{z}_3$ for an arbitrary $\mathbf{C}$. It is easy to see that $\mathbf{z}_3^* = \mathbf{C}^{-1/2}z_3$ has distribution $H_{\mathbf{I},\mathbf{V}}$ and therefore

$$w = \mathbf{b}'\mathbf{C}^{1/2}\mathbf{z}_3^* = ||\mathbf{C}^{1/2}\mathbf{b}||\mathbf{d}'\mathbf{z}_3^*$$

where $||\mathbf{d}|| = 1$. Then the distribution function of w is $G_{\mathbf{V}}(w/||C^{1/2}\mathbf{b}||)$.

To obtain the distribution $G_{\mathbf{V}}$ we can generate a very large sample of $(\mathbf{z}_1, \mathbf{z}_2)$ (say of size 10^6) from $H_{\mathbf{I},\mathbf{V}}$ and use the transformation (13) to generate a sample of $\mathbf{z}_3$ with distribution $G_{\mathbf{V}}$. In this way we can obtain estimates of the quantiles of $G_{\mathbf{V}}$ that can be used for asymptotic inference on any linear combination of the proposed estimator $\widehat{\beta}$. To this end, the matrix $\mathbf{V}$ can be estimated through (12), replacing F by the residual empirical distribution.

This large-sample Monte Carlo can also be used to compute the asymptotic efficiencies of $\widehat{\beta}$ for different error distributions F. We compute the of $\widehat{\beta}$ with respect to the LS estimator (eff_{LS}) and respect to the MM- estimator (eff_{MM}), defined by

$$\text{eff}_{\text{LS}} = \frac{\text{E}(\mathbf{z}_1'\mathbf{C}\mathbf{z}_1)}{\text{E}(\mathbf{z}_3'\mathbf{C}\mathbf{z}_3)},\ \text{eff}_{\text{MM}} = \frac{\text{E}(\mathbf{z}_2'\mathbf{C}\mathbf{z}_2)}{\text{E}(\mathbf{z}_3'\mathbf{C}\mathbf{z}_3)}$$

Since $\mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{z}_3$ are spheric when $\mathbf{C} = \mathbf{I}$, these efficiencies do not depend on $\mathbf{C}$. We compute these efficiencies when F is normal, Student t with 3 and 5 degrees of freedom, and uniform. For p we chose the values $5, 10$ and 20 The results are shown in Table 4. Finally using the same sample we also compute the probabilities that $\widehat{\beta}$ coincides with $\widehat{\beta}_{LS}$ The results are shown in Table 4

	$\mathrm{eff}_{\mathrm{LS}}$			$\mathrm{eff}_{\mathrm{MM}}$			$\mathrm{eff}_{\mathrm{ML-t}}$		
	$p=5$	10	20	5	10	20	5	10	20
Normal	0.998	0.9997	0.9999	1.18	1.18	1.18			
t_3	1.84	1.84	1.84	0.97	0.97	0.97	0.92	0.92	0.92
t_5	1.19	1.19	1.19	1.01	1.01	1.01	0.95	0.95	0.95
Uniform	1.00	1.00	1.00	1.07	1.07	1.07			

Table 4: Asymptotic efficiency of the proposed estimator for four error distributions

$p=$	5	10	20
normal	0.85	0.91	0.96
t_3	0.02	0.001	0.00
t_5	0.14	0.05	0.01
uniform	1.00	1.00	1.00

Table 5: Probability of equality of DCML and LS estimators

Finally using the same sample we also computed the probabilities that $\widehat{\beta}$ coincides with $\widehat{\beta}_{LS}$ The results are shown in Table 5

2.3 Breakdown point

It will be shown that for the estimators employed in this paper, the finite-sample replacement breakdown point of the DCML estimator $\widehat{\beta}$ is that of the initial estimator $\widehat{\beta}_0$.

Consider a data set $\mathbf{Z} = \{\mathbf{z}_i, \, i = 1, ..., n\}$ with $\mathbf{z}_i = (\mathbf{x}_i, y_i)$. Let m be such that $\varepsilon = m/n$ is less than the breakdown point ε^* of $\widehat{\beta}_0$. Let S (the "outlier set") be any set of size m contained in $\{1, ..., n\}$. Let $\mathbf{Z}^* = \{\mathbf{z}_i^*, \; i = 1, .., n\}$ where $\mathbf{z}_i^* = \mathbf{z}_i$ for $i \notin S$ and is arbitrary for $i \in S$. We have to prove that $\widehat{\beta}$ is bounded as a function of $\mathbf{Z}^*$. The following assumptions will be needed.

A) The initial scale $\widehat{\sigma}_0$ is a scale M estimator of the form

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\widehat{\beta}_0}{\widehat{\sigma}_0}\right) = \gamma,$$

where ρ is a "bounded ρ-function" in the sense of (Maronna et al, 2006, p 31), i.e., $\rho \in [0, 1]$, $\rho(0) = 0$, and $\rho(t)$ is a nondecreasing function of $|t|$, which is strictly increasing for $t > 0$ such that $\rho(t) < 1$.

B) The breakdown point of $\widehat{\sigma}_0$ is $\geq \varepsilon^*$.

C) The weight function $W(t)$ in (3) is a nondecreasing function of $|t|$ which is "matched" to ρ in the sense that $W(t) = 0$ implies $\rho(t) = 1$. This is the case in the situations considered here, where $\rho(t) = \rho_{\text{bis}}(t/\mathbf{c}_0.)$ (see (9)-(10)) and $W(t) = \rho'_{\text{bis}}(t/\mathbf{c}_1.)/t$, where $\mathbf{c}_1. > \mathbf{c}_0.$ is chosen to control the efficiency of the MM estimator.

D) Finally we assume

$$n(1 - \varepsilon^* - \gamma) \geq p \tag{14}$$

with γ in (10).

Call h the maximum number of $\mathbf{x}_i$s in a subspace. The maximal breakdown point for $\widehat{\beta}_0$ and $\widehat{\sigma}_0$ is: $\varepsilon^*_{\max} = 0.5(n - h - 1)/n$. Here we have $\gamma = 0.5(n-p)/n \leq \varepsilon^*_{\max}$ since $h \geq p - 1$, which implies (14) since $\varepsilon \leq \varepsilon^*_{\max}$.

We now proceed to the proof. Recall that $\widehat{\beta}$ satisfies

$$\frac{1}{\widehat{\sigma}_0^2}\left(\widehat{\beta} - \widehat{\beta}_0\right)' \mathbf{C}_w \left(\widehat{\beta} - \widehat{\beta}_0\right) \leq \delta,$$

where $\mathbf{C}_w$ is defined in (4). Recall that $\widehat{\beta}_0$, $\widehat{\sigma}_0$ and $\mathbf{C}_w$ depend on $\mathbf{Z}^*$. Since $\varepsilon < \varepsilon^*$ there exist constants a, b, c such that for all S and $\mathbf{Z}^*$:

$$0 < a \leq \widehat{\sigma}_0 \leq b, \quad \left\|\widehat{\beta}_0\right\| \leq c.$$

Also, since $\varepsilon < \varepsilon^*$ there exists $\eta \in (0,1)$ such that

$$n\left(1 - \varepsilon - \frac{\gamma}{1-\eta}\right) \geq p. \tag{15}$$

Let $t_0 > 0$ be such that $\rho\left(t_0\right) = 1 - \eta$, and put $w_0 = W\left(t_0\right)$. Then by (C) $|t| \leq t_0$ implies $W\left(t\right) \geq w_0 > 0$. Let

$$N = N\left(\mathbf{Z}^*\right) = \#\left\{i \notin S : \rho\left(\frac{y_i - \mathbf{x}_i'\widehat{\beta}_0}{\widehat{\sigma}_0}\right) \leq 1 - \eta\right\}.$$

Then it follows from (10) that

$$n\delta \geq \sum_{i \notin S} \rho\left(\frac{y_i - \mathbf{x}_i'\widehat{\beta}_0}{\widehat{\sigma}_0}\right) \geq \left(n - m - N\right)\left(1 - \eta\right),$$

and therefore by (15), since $\varepsilon < \varepsilon^*$

$$N\left(\mathbf{Z}^*\right) \geq n - n\varepsilon - \frac{n\gamma}{1-\eta} \geq p \forall\ \mathbf{Z}^*.$$

Call $\mathcal{A}$ the set of all subsets of $\{1, ..., n\}$ of size $h+1$. Put

$$\lambda_0 = \min_{A \in \mathcal{A}} \lambda_{\min}\left(\sum_{i \in A} \mathbf{x}_i\mathbf{x}_i'\right),$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of a matrix. Then $\lambda_0 > 0$. For any vector $\mathbf{a}$ and all $\mathbf{Z}^*$ we have

$$\mathbf{a}'\mathbf{C}_w\mathbf{a} \geq \mathbf{a}'\left[\sum_{i \notin S} W\left(\frac{y_i - \mathbf{x}_i'\widehat{\beta}_0}{\widehat{\sigma}_0}\right)\mathbf{x}_i\mathbf{x}_i'\right]\mathbf{a} \geq w_0\lambda_0\left\|\mathbf{a}\right\|^2,$$

and therefore we have for all $\mathbf{Z}^*$

$$\delta\widehat{\sigma}_0^2 \geq \left(\widehat{\beta} - \widehat{\beta}_0\right)'\mathbf{C}_w\left(\widehat{\beta} - \widehat{\beta}_0\right) \geq w_0\lambda_0\left\|\widehat{\beta} - \widehat{\beta}_0\right\|^2,$$

which, in view of the boundedness of $\widehat{\beta}_0$ and $\widehat{\sigma}_0$, implies that $\widehat{\beta}$ is bounded.

3 Multivariate estimation

Consider observations $\mathbf{x}_i$, $i = 1, ..., n$ with a normal p-variate distribution $\mathrm{N}_p(\mu, \mathbf{\Sigma})$. Let $\left(\widehat{\mu}_0, \widehat{\mathbf{\Sigma}}_0\right)$ be a robust estimator of multivariate location and scatter. We shall treat μ and $\mathbf{\Sigma}$ separately.

For the estimation of $\mathbf{\Sigma}$ we have, considering μ as a nuisance parameter:

$$d_{\mathrm{KL}}(\mathbf{\Sigma}_0, \mathbf{\Sigma}) = \log|\mathbf{\Sigma}| - \log|\mathbf{\Sigma}_0| + \mathrm{trace}\left(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_0\right) - p, \tag{16}$$

where $|.|$ denotes the determinant. Our procedure amounts to

$$\widehat{\mathbf{\Sigma}} = \arg\min_{\mathbf{\Sigma}}\left[n\log|\mathbf{\Sigma}| + \sum_{i=1}^{n}(\mathbf{x}_i - \mu)\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mu)\right] \tag{17}$$

with $d_{\mathrm{KL}}\left(\widehat{\mathbf{\Sigma}}_0, \mathbf{\Sigma}\right) \leq \delta$.

Call $\widehat{\mathbf{\Sigma}}_{\mathrm{ML}}$ the MLE of $\mathbf{\Sigma}$, i.e. the sample covariance matrix. Put $d_0 = d_{\mathrm{KL}}\left(\widehat{\mathbf{\Sigma}}_0, \widehat{\mathbf{\Sigma}}_{\mathrm{ML}}\right)$. Then using Lagrange multipliers, a straightforward calculation shows that

$$\widehat{\mathbf{\Sigma}} = (1-t)\,\widehat{\mathbf{\Sigma}}_{\mathrm{ML}} + t\widehat{\mathbf{\Sigma}}_0, \tag{18}$$

where $t = 0$ if $d_0 \leq \delta$, and is otherwise determined from the equation $d_{\mathrm{KL}}\left(\widehat{\mathbf{\Sigma}}_0, \mathbf{\Sigma}\right) = \delta$, which is easily derived from (16)-(18).

We now turn to μ. We have

$$d_{\mathrm{KL}}(\mu_0, \mu) = (\mu - \mu_0)'\,\Sigma^{-1}(\mu - \mu_0).$$

The estimator is then defined by

$$\sum_{i=1}^{n}(\mathbf{x}_i - \mu)\,\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mu) = \min \tag{19}$$

with $d_{\mathrm{KL}}(\mu_0, \mu) \leq \delta$. Let $\overline{\mathbf{x}}$ be the sample mean, and define

$$d_0 = (\overline{\mathbf{x}} - \widehat{\mu}_0)'\,\widehat{\Sigma}_0^{-1}(\mathbf{x} - \widehat{\mu}_0).$$

Then a straightforward calculation shows that

$$\widehat{\mu} = t\overline{\mathbf{x}} + (1-t)\,\widehat{\mu}_0 \tag{20}$$

with

$$t = \min\left(1, \sqrt{\frac{\delta}{d_0}}\right).$$

It is easy to show that if the initial estimators are affine-equivariant, so are the resulting ones.

Remark: Unlike the regression and location cases, $d_{\mathrm{KL}}(\mathbf{\Sigma}_0, \mathbf{\Sigma})$ is not symmetric in its arguments. Here we have chosen the form (16) because it yields the simple intuitive result (18), while the alternative order yields a more complicated result.

3.1 Simulations

As initial estimator we employ an S estimator (Davies, 1987) with bisquare scale, computed as described at the end of page 199 of (Maronna et al, 2006). It is implemented as the function CovSest with the option method= "bisquare" in the R package rrcov.

This study includes $p = 2$, 5 and 10. The reason why larger values of p are not included is the following. Rocke (1996) found out that the efficiency of S estimators with a monotone weight function increases with p, and therefore there is little to be gained with DCML when p is large.

We now define the S estimator. For $(\mu, \mathbf{\Sigma})$ denote the (squared) Mahalanobis distance of $\mathbf{x}$ as

$$d(\mathbf{x}, \mu, \mathbf{\Sigma}) = (\mathbf{x} - \mu)' \, \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu).$$

	a	b	c
$\mathbf{\Sigma}$	1.02	0.82	0.18
μ	0.55	0.88	-0.30

Table 6: Constants for the approximate computation of δ

Define a scale M estimator $\widehat{\sigma} = \widehat{\sigma}\left(\mu, \mathbf{\Sigma}\right)$ as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{d\left(\mathbf{x}, \mu, \mathbf{\Sigma}\right)^{1/2}}{\sigma}\right) = \gamma,$$

where ρ is the bisquare ρ-function (9), and $\gamma = 0.5\left(1 - p/n\right)$ which ensures maximal breakdown point. The S estimator is defined by

$$\left(\widehat{\mu}_0, \widetilde{\mathbf{\Sigma}}\right) = \arg\min\left\{\widehat{\sigma}\left(\mathbf{t}, \mathbf{V}\right) : \mathbf{t} \in R^p, |\mathbf{V}| = 1\right\}$$

Since $|\widetilde{\mathbf{\Sigma}}| = 1$, we have to scale $\widetilde{\mathbf{\Sigma}}$ to make it a consistent estimator of the covariance matrix under normality. Put $d_i = d\left(\mathbf{x}_i, \widehat{\mu}_0, \widetilde{\mathbf{\Sigma}}\right)$ and call χ^2_p the chi-squared distribution with p degrees of freedom. Then define

$$\widehat{\mathbf{\Sigma}}_0 = \frac{\text{median}_i\{d_i\}}{\text{median}(\chi^2_p)}\widetilde{\mathbf{\Sigma}}.$$

The constants δ in (17) and (19) were chosen as

$$\delta = an^{-b}p^c, \tag{21}$$

with (a, b, c) given in Table 6.

The motivation for this choice is as follows. It was considered as reasonable to choose for each (p, n), δ as some $\alpha-$quantile of d_{KL} under the nominal model, i.e. the multivariate normal distribution. Exploratory simulations suggested α between 0.4 and 0.6. The quantiles were computed by simulation for p between 2 and 10 and n between $5p$ and 500. Then for each α the α-quantile was fitted by regression as a function of n and p of the form (21). Finally, after the simulation was completed, it was decided that $\alpha = 0.4$ yielded the best results.

The values of c indicate that when p increases, the quantiles for $\mathbf{\Sigma}$ increase very slowly, and those for μ decrease. This fact may seem counter-intuitive, but it is a consequence of the increasing efficiency of the S estimator: when p increases, the S estimator becomes "closer" to the classical one, which makes d_{KL} smaller.

For each n and p we generate N_{rep} samples of size n from $\mathrm{N}_p\left(\mathbf{0}, \mathbf{I}\right)$. For a contamination rate ε, the first $m = [n\varepsilon]$ elements are replaced by $(K, 0, ..., 0)$ where K ranges between 1 and 10. For each sample three estimators were computed: the sample mean and covariance matrix, the S estimator, and the DCML estimator given by (18)-(20).

For each scenario, each estimator is evaluated by its "loss" defined as $\|\mu\|^2$ for location and as $d_{\mathrm{KL}}\left(\mathbf{I}, \mathbf{\Sigma}\right) = \mathrm{trace}\left(\mathbf{\Sigma}\right) - \log\left|\mathbf{\Sigma}\right|$ for scatter and the results were summarized by the respective mean losses. Table 7 shows the efficiencies, defined as the ratio of the mean losses of the classical and the robust estimator.

It is seen that DCML is able to substantially increase the efficiency of S-E, especially for $p = 2$. The efficiency for location is much higher than for scatter. The fact that the efficiency of S-E increases with p is also clear. Actually, for $p = 15$ the efficiency of S-E is ≥ 0.96.

Table 8 shows the maximum mean losses for contamination rate $\varepsilon = 0.1$. It is seen that in general the price for the increase in efficiency is at worst a small

		$\mathbf{\Sigma}$		μ	
p	n	S-E	DCML	S-E	DCML
2	10	0.422	0.627	0.690	0.889
	20	0.414	0.692	0.673	0.898
	40	0.407	0.762	0.586	0.867
5	25	0.772	0.922	0.893	0.971
	50	0.778	0.962	0.876	0.980
	100	0.777	0.977	0.855	0.978
10	50	0.936	0.994	0.955	0.996
	100	0.921	0.995	0.946	0.995
	200	0.914	0.996	0.945	0.998

Table 7: Efficiencies of estimators

increase of the maximum loss and at best a decrease thereof. Figure 2 compares the losses of S-E and DCML as a function of the outlier size K for $\varepsilon = 0.1$.

3.2 Breakdown point

It is easy to show that the replacement breakdown point of the DCML estimators is that of the initial ones. We give the details for $\widehat{\mathbf{\Sigma}}$, the case of $\widehat{\mu}$ being similar. Consider a data set $\mathbf{X} = \{\mathbf{x}_i \; i = 1, ..., n\}$. Let m be such that $\varepsilon = m/n$ is less than the breakdown point ε^* of the initial estimator $\widehat{\mathbf{\Sigma}}_0$. Let $\mathbf{X}^*$ be a data set that coincides with $\mathbf{X}$ except for m elements which are arbitrary. We have to prove that, as a function of $\mathbf{X}^*$, the largest eigenvalue $\lambda_{\max}$ of $\widehat{\mathbf{\Sigma}}$ is bounded, and the smallest one $\lambda_{\min}$ is bounded away from zero. We know that this property holds

			$\mathbf{\Sigma}$		μ	
ε	p	n	S-E	DCML	S-E	DCML
0.1	2	10	0.91	1.03	0.32	0.34
		20	0.51	0.53	0.18	0.20
		40	0.27	0.31	0.10	0.11
	5	25	1.01	1.03	0.26	0.27
		50	0.65	0.72	0.17	0.20
		100	0.39	0.48	0.11	0.13
	10	50	2.90	3.26	0.44	0.51
		100	1.82	2.06	0.28	0.31
		200	1.39	1.71	0.21	0.27
0.2	2	10	1.42	1.49	0.50	0.51
		20	0.95	0.77	0.34	0.37
		40	0.68	0.50	0.28	0.27
	5	25	3.60	3.43	0.97	1.19
		50	2.49	3.39	0.67	0.87
		100	2.20	2.13	0.56	0.73
	10	50	11.21	11.15	2.49	2.88
		100	6.46	6.50	1.65	1.95
		200	5.71	5.72	1.52	1.78

Table 8: Simulation: maximum mean losses of estimators

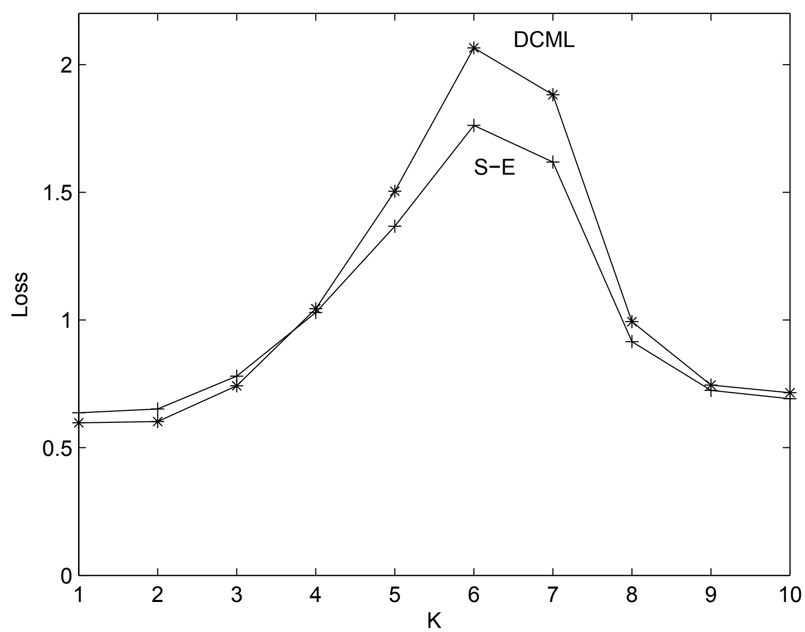

Figure 2: Losses of scatter matrices for $p = 10$, $n = 100$ and 10% contamination, as a function of outlier size.

for $\widehat{\boldsymbol{\Sigma}}_0$. Since by (16)

$$\log|\widehat{\boldsymbol{\Sigma}}| - \log|\widehat{\boldsymbol{\Sigma}}_0| + \text{trace}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}_0\right) - p \leq \delta,$$

it follows from the "trace" term that $\lambda_{\min}$ cannot tend to zero, and then it follows from the "log" term that $\lambda_{\max}$ cannot tend to infinity. ■

4 Real data

In this section we apply the estimators to two published data sets.

Computed with	LS	S-E	MM	G-Y	B-S	DCML
Good data	1.095	1.416	1.126	1.095	1.095	1.095
Whole data	1.921	1.143	1.100	1.322	1.484	1.164

Table 9: Stack los data: prediction RMSEs of estimators for "good" data

4.1 Regression

We consider the well-known stack loss data set with $n = 21$ and $p = 3$ plus intercept. Lacking a "true model" we have to employ alternative criteria for robustness and efficiency.

There seems to be a general agreement to consider observations 1, 3, 4 and 21 as atypical; see (Rousseew and Leroy, 1987). Call "good data" the data set without {1,3,4,21}. The estimators were first computed using the good data, and the root mean squared prediction errors (RMSE: square root of the mean of the squared residuals) was computed for the same data. The comparison with LS was employed as a surrogate criterion for efficiency. For a surrogate criterion for robustness, the estimators were then computed for the whole data set, and the RMSE again computed *only* for the good data. Table 9 shows the results.

The first row shows that G-Y, B-S and DCML are here "fully efficient", S-E is rather inefficient, and MM has a high efficiency. The second row shows S-E, MM and DCML as most robust, followed by G-Y, and B-S as the less robust one.

The behavior of S-E is puzzling. It gives zero weights to some "good" observations. The estimator was recomputed several times to rule out the effect of the subsampling.

4.2 Multivariate estimation

Here we choose the Philips Mecoma data, employed in Problem 1 in (Rousseeuw and Van Driessen, 1999), with $n = 677$ and $p = 9$. Plotting the Mahalanobis distances from the S estimator shows a number of clear outliers, the sequence with indexes between 491 and 565 being the most outstanding ones. We defined as "bad data" the observations with Mahalanobis distances larger than 60, which yielded 80 observations. Lacking a criterion similar to prediction error like in the former example, we defined as the "true parameters" the MLE (mean and covariance matrix) applied to the "good" data, which will be called μ_{good} and $\mathbf{\Sigma}_{\text{good}}$, respectively.

We then computed, as above, the estimators based on the "good" data and their Kullback-Leibler distances to the "truth"; and then did the same for the estimators based on the whole data. Namely, we computed

$$d = \text{trace}\left(\mathbf{\Sigma}_{\text{god}}^{-1}\mathbf{V}\right) - p - \log\left|\mathbf{\Sigma}_{\text{good}}^{-1}\mathbf{V}\right|$$

for each scatter estimator $\mathbf{V}$, and

$$d = \left(\mathbf{t} - \mu_{\text{good}}\right)' \mathbf{C}_{\text{good}}^{-1}\left(\mathbf{t} - \mu_{\text{good}}\right)$$

for each location estimator $\mathbf{t}$. Table 10 shows the results.

It is seen that here DCML outperforms S-E in all cases.

References

Bondell, H.D. and Stefanski, L.A. (2013). "Efficient Robust Regression via Two-Stage Generalized Empirical Likelihood", *Journal of the American Statistical Association,* 108, 644-655

	Computed with	MLE	S-E	DCML
Scatter	Good data		0.381	0.286
	Whole data	6.282	0.381	0.322
Location	Good data		0.051	0.039
	Whole data	1.067	0.051	0.044

Table 10: Philips data: Kullback-Leibler distances between estimators and “true values”

Davies, P.L. (1987), “Asymptotic Behavior of S-estimators of Multivariate Location Parameters and Dispersion Matrices”, *The Annals of Statistics*, 15, 1269–1292.

Gervini, D. and Yohai, V.J. (2002), “A Class of Robust and Fully Efficient Regression Estimators”, *The Annals of Statistics*, 30, 583-616.

Hampel, F.R. (1971), “A General Definition of Qualitative Robustness”, *The Annals of Mathematical Statistics*, 42, 1887–1896.

Huber, P.J. (1964), “Robust Estimation of a Location Parameter”, *The Annals of Mathematical Statistics,* 35**,** 73-101.

Maronna, R.A., Martin, R.D. and Yohai, V.J.(2006). *Robust Statistics: Theory and Methods*, New York: John Wiley and Sons.

Rocke, D.M. (1996), “Robustness Properties of S-estimators of Multivariate Location and Shape in High Dimension”, *The Annals of Statistics,* 24**,** 1327-1345.

Rousseeuw, P.J and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley and Sons.

Rousseeuw, P.J. and Van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator”, *Technometrics* 41, 212-223.

Rousseeuw, P.J and Yohai, V.J. (1984), “Robust Regression by Means of S–

estimators", In *Robust and Nonlinear Time Series*, J. Franke, W. Härdle y R. D. Martin (eds.). Lectures Notes in Statistics, 26, 256–272, New York: Springer Verlag.

Yohai, V.J. (1987), "High Breakdown–point and High Efficiency Estimates for Regression", *The Annals of Statistics,* 15, 642–65.

Yohai, V.J., Stahel, W.A. and Zamar, R.H. (1991), "A Procedure for Robust Estimation and Inference in Linear Regression", In *Directions in Robust Statistics and Diagnostics (Part II),* W. Stahel and S. Weisberg, eds., The IMA Volumes in Mathematics and is Applications, New York: Springer Verlag, 365-374.

Yohai, V.J. and Zamar, R.H. (1988), "High Breakdown Estimates of Regression by Means of the Minimization of an Efficient Scale", *Journal of the American Statistical Association,* 83, 406–413.